

# A Simple Converse of Burnashev's Reliability Function

Peter Berlin<sup>†</sup>, Barış Nakiboğlu<sup>‡</sup>, Bixio Rimoldi<sup>†</sup>, Emre Telatar<sup>†</sup>

<sup>†</sup>School of Computer and Communication Sciences  
Ecole Polytechnique Fédérale de Lausanne (EPFL)  
CH-1015 Lausanne, Switzerland

<sup>‡</sup> Department of Electrical Engineering and Computer Science  
Massachusetts Institute of Technology (MIT)  
MA 02139 Cambridge, USA

peter.berlin@epfl.ch, nakib@mit.edu,  
bixio.rimoldi@epfl.ch, emre.telatar@epfl.ch

## Abstract

In a remarkable paper published in 1976, Burnashev determined the reliability function of variable-length block codes over discrete memoryless channels with feedback. Subsequently, an alternative *achievability* proof was obtained by Yamamoto and Itoh via a particularly simple and instructive scheme. Their idea is to alternate between a communication and a confirmation phase until the receiver detects the codeword used by the sender to acknowledge that the message is correct. We provide a *converse* that parallels the Yamamoto-Itoh achievability construction. Besides being simpler than the original, the proposed converse suggests that a communication and a confirmation phase are implicit in any scheme for which the probability of error decreases with the largest possible exponent. The proposed converse also makes it intuitively clear why the terms that appear in Burnashev's exponent are necessary.

## Index Terms

Burnashev's error exponent, discrete memoryless channels (DMCs), feedback, variable-length communication

## I. INTRODUCTION

It is well known (see e.g. [1] and [2]), that the capacity of a discrete memoryless channel (DMC) is not increased by feedback.<sup>1</sup> Nevertheless, feedback can help in at least two ways: for a fixed target error probability, feedback can be used to reduce the sender/receiver complexity and/or to reduce the expected decoding delay. An example is the binary erasure channel, where feedback makes it possible to implement a communication strategy that is extremely simple and also minimizes the delay. The strategy is simply to send each information bit repeatedly until it is received unerased. This strategy is capacity achieving, results in zero probability of error, and reproduces each information bit with the smallest delay among all possible strategies.

The reliability function—also called the error exponent—is a natural way to quantify the benefit of feedback. For block codes on channels without feedback the reliability function is defined as

$$E(R) = \limsup_{T \rightarrow \infty} -\frac{1}{T} \ln P_e(\lceil e^{RT} \rceil, T), \quad (1)$$

where  $P_e(M, T)$  is the smallest possible error probability of length  $T$  block codes with  $M$  codewords.

<sup>1</sup>According to common practice, we say that feedback is available if the encoder may select the current channel input as a function not only of the message but also of all past channel outputs.

The decoding time  $T$  in a communication system with feedback may depend on the channel output sequence.<sup>2</sup> If it does, the decoding time  $T$  becomes a random variable and the notions of rate and reliability function need to be redefined. Following Burnashev [3], in this case we define the rate as

$$R \triangleq \frac{\ln M}{E[T]}, \quad (2)$$

where  $M$  is the size of the message set. Similarly we define the reliability function as

$$E_f(R) \triangleq \lim_{t \rightarrow \infty} -\frac{1}{t} \ln P_{e,f}(\lceil e^{Rt} \rceil, t), \quad (3)$$

where  $P_{e,f}(M, t)$  is the smallest error probability of a variable-length block code with feedback that transmits one of  $M$  equiprobable messages by means of  $t$  or fewer channel uses on average. As we remark below, the limit exists for all rates from zero to capacity.

Burnashev showed that for a DMC of capacity  $C$ , the reliability function  $E_f(R)$  equals

$$E_B(R) = C_1(1 - R/C), \quad 0 \leq R \leq C, \quad (4)$$

where  $C_1$  is determined by the two “most distinguishable” channel input symbols as

$$C_1 = \max_{x, x'} D(p(\cdot|x) \| p(\cdot|x')),$$

where  $p(\cdot|x)$  is the probability distribution of the channel output when the input is  $x$ , and  $D(\cdot\|\cdot)$  denotes the Kullback-Liebler divergence between two probability distributions. It is remarkable that (4) determines the reliability function exactly for all rates. In contrast, the reliability function without feedback is known exactly only for rates above a critical rate. Below the critical rate only upper and lower bounds to the reliability function without feedback are known. For a binary symmetric channel the situation is depicted in Fig. 1.

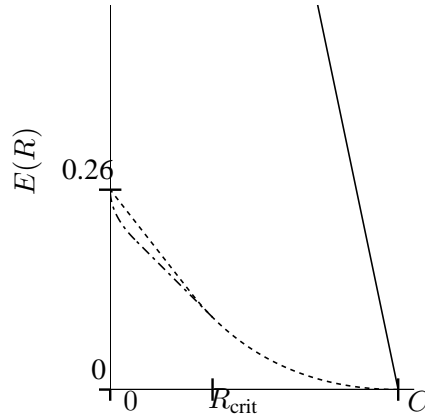


Fig. 1. Reliability functions for a binary symmetric channel with crossover probability 0.1. Shown is Burnashev's reliability function for channels with feedback (solid line) and upper and lower bounds to the reliability function for channels without feedback. The upper bound (dotted line) is given by the *straight line bound* at low rates and by the *sphere packing bound* at higher rates. The lower bound (dot-dashed line) is given by the *expurgated bound*. The upper and lower bounds coincide above the critical rate,  $R_{\text{crit}}$ .

Burnashev showed that  $E_f = E_B$  by showing that for every communication scheme

$$E[T] \geq \left( \frac{\ln M}{C} - \frac{\ln P_e}{C_1} \right) (1 - o(1)) \quad (5)$$

<sup>2</sup> If the decoding time is not fixed, in the absence of feedback the sender may not know when the receiver has decoded. This problem does not exist if there is feedback.

where  $o(1)$  represents positive terms that tend to zero as  $1/P_e$  tends to infinity, and that there exists schemes with

$$E[T] \leq \left( \frac{\ln M}{C} - \frac{\ln P_e}{C_1} \right) (1 + o(1)), \quad (6)$$

where  $o(1)$  now represents positive terms that tend to zero as both  $M$  and  $1/P_e$  tend to infinity.

For a plausibility argument that justifies (6) it suffices to summarize the achievability construction by Yamamoto and Itoh [4]. Their scheme relies on two distinct transmission phases that we shall call the communication and the confirmation phase, respectively. In the communication phase the message is encoded using a fixed-length block code and the codeword is transmitted over the forward channel. The decoder makes a tentative decision based on the corresponding channel output. The encoder knows the channel output and can run the algorithm used by the receiver to determine the tentative decision. If the tentative decision is correct, in the confirmation phase the encoder sends ACK. Otherwise it sends NACK. ACKs and NACKs are sent via a fixed-length repetition code. (The code consists of two codewords). During the confirmation phase the decoder performs a binary hypothesis test to decide if ACK or NACK was transmitted. If ACK is decoded, the tentative decision becomes final and the transmission of the current message ends, leaving the system free to restart with a new message. If NACK is decoded, the tentative decision is discarded and the two phase scheme restarts with the same message.

The overhead caused by retransmissions is negligible if the probability of decoding NACK is small. This is the case if both the error probability of the communication phase as well as that of the confirmation phase are small. Assuming that this is the case, the number of channel uses for the communication phase (including repetitions) is slightly above  $(\ln M)/C$ . The probability of error is the probability that NACK is sent and ACK is decoded. In the asymptotic regime of interest this probability is dominated by the probability that ACK is decoded given that NACK is sent. In a straightforward application of Stein's lemma [5] one immediately sees that we can make this probability to be slightly less than  $P_e$  (thus achieve error probability  $P_e$ ) by means of a confirmation code of length slightly above  $(-\ln P_e)/C_1$ . Summing up, we see that we can make the error probability arbitrarily close to  $P_e$  by means of slightly more than  $(\ln M)/C - (\ln P_e)/C_1$  channel uses on average. This confirms (6).

To obtain the converse (5), Burnashev investigated the entropy of the *a posteriori* probability distribution over the message set. He showed that the average decrease of this entropy due to an additional channel output observation, as well as the average decrease of the logarithm of this entropy, are bounded. He uses these bounds to form two submartingales, one based on the entropy of the *a posteriori* distribution and the other based on the logarithm of this entropy. He then constructs a single submartingale by patching these two together. Then Doob's optional stopping theorem is applied to this submartingale and the desired bound on the expected decoding time, which is a stopping time, is obtained. Burnashev's proof is an excellent example of the power of martingales, however both the sophistication of the martingale construction and the use of the logarithm of entropy leaves the reader with little insight about some of the terms in the converse bound. While it is easy to see that  $(\ln M)/C$  channel uses are needed in average, it was not as clear why one needs an additional  $(-\ln P_e)/C_1$  channel uses. The connection of the latter term to binary hypothesis testing suggested the existence of an operational justification. The work presented in this paper started as an attempt to find this operational justification.

Our converse somewhat parallels the Yamamoto–Itoh achievability scheme. This suggests that that a communication and confirmation phase may be implicit components of any scheme for which the probability of error decreases with the largest possible exponent. Our approach has been generalized by Como, Yüksel and Tatikonda in [6] to prove a similar converse for variable-length block codes on Finite State Markov Channels.

## II. CHANNEL MODEL AND VARIABLE-LENGTH CODES AS TREES

We consider a discrete memoryless channel, with finite input alphabet  $\mathcal{X}$ , finite output alphabet  $\mathcal{Y}$ , and transition probabilities  $p(y|x)$ . We will denote the channel input and output symbols at time  $n$  by  $X_n$  and  $Y_n$ , and denote the corresponding vectors  $(X_1, X_2, \dots, X_n)$  and  $(Y_1, Y_2, \dots, Y_n)$  by  $X^n$  and  $Y^n$ , respectively. A perfect causal

feedback link is available, i.e., at time  $n$  the encoder knows  $y^{n-1}$ . (Following common practice, random variables are represented by capital letters and their realizations are denoted by the corresponding lowercase letters.)

We will assume, without loss of generality, that the channel has no “useless outputs symbols”, i.e., no symbols  $y$  for which  $p(y|x) = 0$  for every  $x$ . Note that for channels for which  $C_1$  is infinite, the lower bound to the expected decoding time is a restatement of the fact that feedback does not increase capacity. We will therefore restrict our attention to channels for which  $C_1 < \infty$ . For such channels,  $p(y|x) > 0$  for every  $x$  and  $y$ ; if not, there exists an  $x$  and  $y$  for which  $p(y|x) = 0$ . Since  $y$  is reachable from some input, there also exists an  $x'$  for which  $p(y|x') > 0$ . But then  $D(p(\cdot|x')||p(\cdot|x)) = \infty$  contradicting the finiteness of  $C_1$ . The fact that both  $\mathcal{X}$  and  $\mathcal{Y}$  are finite sets lets us further conclude that for the channels of interest to this paper, there is a  $\lambda > 0$  for which  $p(y|x) \geq \lambda$  for every  $x$  and  $y$ .

A variable-length block code is defined by two maps: the encoder and the decoder. The encoder<sup>3</sup> functions  $f_n(\cdot, \cdot) : \mathcal{W} \times \mathcal{Y}^{n-1} \rightarrow \mathcal{X}$ , where  $\mathcal{W} = \{1, \dots, M\}$  is the set of all possible messages, determine the channel input  $X_n = f_n(W, Y^{n-1})$  based on the message  $W$  and on past channel outputs  $Y^{n-1}$ . The decoder function  $\hat{W}(\cdot) : \mathcal{Z} \rightarrow \mathcal{W}$ , where  $\mathcal{Z}$  is the receiver observation space until the decoding time  $T$ , i.e.,  $Y^T$  takes values in  $\mathcal{Z}$ . The decoding time  $T$  should be a stopping time<sup>4</sup> with respect to the receiver observation  $Y^n$  otherwise the decision of when to decode would depend on future channel outputs and the decoder would no longer be causal. We treat the case when  $E[T] < \infty$ , and point out that what we are setting out to prove, namely (5), is trivially true when  $E[T] = \infty$ .

The codes we consider here differ from non-block (also called sequential) codes with variable delay, such as those studied in [8] and [9]. In sequential coding, the message (typically as an infinite stream of bits) is introduced to the transmitter and decoded by the receiver in a progressive fashion. Delay is measured separately for each bit, and is defined as the time between the introduction and decoding of the bit. This is in contrast to the codes considered in this paper, where the entire message is introduced to the transmitter at the start of communication, and  $T$  measures the duration of the communication. Due to their different problem formulation, sequential codes with feedback have reliability functions that differ from those for variable-length block codes, just as fixed constraint length convolutional codes have reliability functions that differ from those of fixed-length block codes.

The observation space  $\mathcal{Z}$  is a collection of channel output sequences and for a DMC with feedback the length of these sequences may vary. (The length of the channel input itself may depend on the channel realization). Nevertheless, these sequences have the property of being prefix-free (otherwise the decision to stop would require knowledge of the future). Thus,  $\mathcal{Z}$  can be represented as the leaves of a complete  $|\mathcal{Y}|$ -ary tree  $\mathcal{T}$  (complete in the sense that each intermediate node has  $|\mathcal{Y}|$  descendants), and has expected depth  $E[T] < \infty$ . Note that the decision time  $T$  is simply the first time the sequence  $Y_1, Y_2, \dots$  of channel outputs hits a leaf of  $\mathcal{T}$ . Furthermore we may label each leaf of  $\mathcal{T}$  with the message decoded by the receiver when that leaf is reached. This way the decoder is completely specified by the labeled tree  $\mathcal{T}$ . The message statistics, the code, and the transition probabilities of the channel determine a probability measure on the tree  $\mathcal{T}$ .

### III. BINARY HYPOTHESIS TESTING WITH FEEDBACK

The binary case ( $M = 2$ ) will play a key role in our main proof. In this section we assume that the message set contains only two elements. We will arbitrarily denote the two hypotheses by  $A$  and  $N$  (ACK and NACK, respectively). We denote by  $\mathcal{Q}_A$  and  $\mathcal{Q}_N$  the corresponding probability distributions on the leaves of  $\mathcal{T}$ .

The following proposition bounds the Kullback-Leibler divergence  $D(\mathcal{Q}_A||\mathcal{Q}_N)$ . It will be used in the main result of this section to bound the error probability of binary hypothesis testing with feedback. The reader familiar

<sup>3</sup> For clarity of exposition we will only treat deterministic coding strategies here. Randomized strategies may be included without significant modification to the core of the proof.

<sup>4</sup> A discussion of stopping times can be found in [7, sect. 10.8].

with Stein's Lemma will not be surprised by the fact that the Kullback-Leibler divergence  $D(\mathcal{Q}_A \parallel \mathcal{Q}_N)$  plays a key role in binary hypothesis testing with feedback. The steps here closely parallel those in [10, Sec. III] and [11, Sec. 2.2].

*Proposition 1:* For any binary hypothesis testing scheme for a channel with feedback

$$D(\mathcal{Q}_A \parallel \mathcal{Q}_N) \leq C_1 E[T | A]$$

where  $T$  is the decision stopping time,  $E[T] < \infty$ , and  $E[T | A]$  denotes the expectation of  $T$  conditioned on hypothesis  $A$ .

*Proof:* In the following, we will denote probability under hypothesis  $A$  by  $P_A(\cdot)$  and probability under hypothesis  $N$  by  $P_N(\cdot)$ . Let

$$V_n = \ln \frac{P_A(Y_1, \dots, Y_n)}{P_N(Y_1, \dots, Y_n)}, \quad (7)$$

so that  $D(\mathcal{Q}_A \parallel \mathcal{Q}_N) = E[V_T | A]$ , and the proposition is equivalent to the statement  $E[V_T - C_1 T | A] \leq 0$ . Observe that

$$V_n = \sum_{k=1}^n U_k \quad \text{where } U_k = \ln \frac{P_A(Y_k | Y^{k-1})}{P_N(Y_k | Y^{k-1})}. \quad (8)$$

Note now that

$$\begin{aligned} E[U_k | A, Y^{k-1}] &= E \left[ \ln \frac{P_A(Y_k | Y^{k-1})}{P_N(Y_k | Y^{k-1})} \middle| A, Y^{k-1} \right] \\ &= E \left[ \ln \frac{P_A(Y_k | X_k = f_k(A, Y^{k-1}), Y^{k-1})}{P_N(Y_k | X_k = f_k(N, Y^{k-1}), Y^{k-1})} \middle| A, X_k = f_k(A, Y^{k-1}), Y^{k-1} \right] \\ &= \sum_{y \in \mathcal{Y}} \Pr \{Y_k = y | X_k = f_k(A, Y^{k-1})\} \ln \frac{\Pr \{Y_k = y | X_k = f_k(A, Y^{k-1})\}}{\Pr \{Y_k = y | X_k = f_k(N, Y^{k-1})\}} \\ &\leq C_1, \end{aligned} \quad (9)$$

where  $f_k(\cdot, \cdot)$  is the encoder function at time  $k$ . Consequently,  $\{V_n - nC_1\}$  is a supermartingale under hypothesis  $A$ . Observe that the existence of a  $\lambda > 0$  for which  $p(y|x) > \lambda$  for all  $x, y$  implies that  $|U_k| < \ln \frac{1}{\lambda}$ . We can now use Doob's Optional-Stopping Theorem (see e.g. [7, Sec. 10.10]) to conclude that  $E[V_T - C_1 T | A] \leq 0$ . ■

We can apply Proposition 1 to find a lower bound on the error probability of a binary hypothesis testing problem with feedback. The bound is expressed in terms of the expected decision time.

*Lemma 1:* The error probability of a binary hypothesis test performed across a DMC with feedback and variable-length codes is lower bounded by

$$P_e \geq \frac{\min\{p_A, p_N\}}{4} e^{-C_1 E[T]}$$

where  $p_A$  and  $p_N$  are the *a priori* probabilities of the hypotheses.

*Proof:* Each decision rule corresponds to a tree where each leaf  $Y^T$  is associated with a decoded hypothesis  $\hat{W}(Y^T)$ . Thus we can partition the leaves into two sets corresponding to the two hypotheses.

$$\begin{aligned} \mathcal{S} &= \{y^T : \hat{W}(y^T) = A\} \\ \bar{\mathcal{S}} &= \{y^T : \hat{W}(y^T) \neq A\} \end{aligned}$$

where  $\mathcal{S}$  is the decision region for hypothesis  $A$ .

The log sum inequality [12], [13] (or data processing lemma for divergence) implies

$$D(\mathcal{Q}_A \parallel \mathcal{Q}_N) \geq \mathcal{Q}_A(\mathcal{S}) \ln \frac{\mathcal{Q}_A(\mathcal{S})}{\mathcal{Q}_N(\mathcal{S})} + \mathcal{Q}_A(\bar{\mathcal{S}}) \ln \frac{\mathcal{Q}_A(\bar{\mathcal{S}})}{\mathcal{Q}_N(\bar{\mathcal{S}})}. \quad (10)$$

By Proposition 1,  $C_1 E[T|A] \geq D(\mathcal{Q}_A \| \mathcal{Q}_N)$ , thus (10) can be re-arranged to give

$$C_1 E[T|A] \geq -\mathcal{Q}_A(\mathcal{S}) \ln \mathcal{Q}_N(\mathcal{S}) - h(\mathcal{Q}_A(\bar{\mathcal{S}})), \quad (11)$$

where  $h(\cdot)$  is the binary entropy function. Writing the overall probability of error in terms of marginal error probabilities yields

$$P_e = p_N \mathcal{Q}_N(\mathcal{S}) + p_A \mathcal{Q}_A(\bar{\mathcal{S}}),$$

which allows us to bound  $\mathcal{Q}_N(\mathcal{S})$  as

$$\mathcal{Q}_N(\mathcal{S}) \leq \frac{P_e}{p_N} \leq \frac{P_e}{\min\{p_A, p_N\}}.$$

Substituting back into (11) yields a bound on the expected depth of the decision tree conditioned on  $A$  just in terms of  $\mathcal{Q}_A$  and the *a priori* message probabilities

$$C_1 E[T|A] \geq -\mathcal{Q}_A(\mathcal{S}) \ln \frac{P_e}{\min\{p_A, p_N\}} - h(\mathcal{Q}_A(\bar{\mathcal{S}})). \quad (12)$$

Following identical steps with the roles of  $A$  and  $N$  swapped yields

$$C_1 E[T|N] \geq -\mathcal{Q}_N(\bar{\mathcal{S}}) \ln \frac{P_e}{\min\{p_A, p_N\}} - h(\mathcal{Q}_N(\mathcal{S})). \quad (13)$$

We can now average both sides of (12) and (13) by weighting with the corresponding *a priori* probabilities. If we do so and use the facts that  $p_A \mathcal{Q}_A(\mathcal{S}) + p_N \mathcal{Q}_N(\bar{\mathcal{S}})$  is the probability of making the correct decision and  $p_A \mathcal{Q}_A(\bar{\mathcal{S}}) + p_N \mathcal{Q}_N(\mathcal{S})$  is the probability of making an error together with the concavity of the binary entropy function, we obtain the following unconditioned bound on the depth of the decision tree

$$\begin{aligned} C_1 E[T] &\geq -(1 - P_e) \ln \frac{P_e}{\min\{p_A, p_N\}} - h(P_e) \\ &\geq -\ln P_e - 2h(P_e) + \ln \min\{p_A, p_N\} \\ &\geq -\ln P_e - 2\ln 2 + \ln \min\{p_A, p_N\}. \end{aligned}$$

Solving for  $P_e$  completes the proof. ■

It is perhaps worthwhile pointing out why the factor  $\min\{p_A, p_N\}$  arises: if one of the hypotheses has small *a priori* probability, one can achieve an equally small error probability by always deciding for the other hypothesis, irrespective of the channel observations.

#### IV. EXPECTED TREE DEPTH AND CHANNEL CAPACITY

Given the channel observations  $y^n$ , one can calculate the *a posteriori* probability  $p_{W|Y^n}(w|y^n)$  of any message  $w \in \mathcal{W}$ . Recall that a maximum *a posteriori* (MAP) decoder asked to decide at time  $n$  when  $Y^n = y^n$  will chose (one of) the message(s) that has the largest *a posteriori* probability  $p_{\max} = \max_w p_{W|Y^n}(w|y^n)$ . The probability of error will then be  $P_e(y^n) = 1 - p_{\max}$ . Similarly, we can define the probability of error of a MAP decoder for each leaf of the observation tree  $\mathcal{T}$ . Let us denote by  $P_e(y^T)$  the probability of error given the observation  $y^T$ . The unconditioned probability of error is then  $P_e = E[P_e(Y^T)]$ .

For any fixed  $\delta > 0$  we can define a stopping time  $\tau$  as the first time that the error probability goes below  $\delta$ , if this happens before  $T$ , and as  $T$  otherwise:

$$\tau = \inf \{n : (P_e(y^n) \leq \delta) \text{ or } (n = T)\} \quad (14)$$

If  $P_e(Y^\tau)$  exceeds  $\delta$ , then we are certain that  $\tau = T$ , and  $P_e(Y^n) > \delta$  for all  $0 \leq n \leq T$ , so the event  $P_e(Y^\tau) > \delta$  is included in the event  $P_e(Y^T) > \delta$ . (We have inclusion instead of equality since  $P_e(Y^\tau) \leq \delta$  does not exclude  $P_e(Y^T) > \delta$ .) Thus

$$\Pr \{P_e(Y^\tau) > \delta\} \leq \Pr \{P_e(Y^T) > \delta\} \leq \frac{P_e}{\delta}, \quad (15)$$

where the second inequality is an application of Markov's inequality.

Given a particular realization  $y^n$  we will denote the entropy of the *a posteriori* distribution  $p_{W|Y^n}(\cdot|y^n)$  as  $\mathcal{H}(W|y^n)$ . Then  $\mathcal{H}(W|Y^n)$  is a random variable<sup>5</sup> and  $E[\mathcal{H}(W|Y^n)] = H(W|Y^n)$ . If  $P_e(y^\tau) \leq \delta \leq \frac{1}{2}$ , then from Fano's inequality it follows that

$$\mathcal{H}(W|y^\tau) \leq h(\delta) + \delta \ln M. \quad (16)$$

The expected value of  $\mathcal{H}(W|Y^\tau)$  can be bounded by conditioning on the event  $P_e(Y^\tau) \leq \delta$  and its complement then applying (16) and then (15) as follows

$$\begin{aligned} E[\mathcal{H}(W|Y^\tau)] &= E[\mathcal{H}(W|Y^\tau) | P_e(Y^\tau) \leq \delta] \Pr \{P_e(Y^\tau) \leq \delta\} + E[\mathcal{H}(W|Y^\tau) | P_e(Y^\tau) > \delta] \Pr \{P_e(Y^\tau) > \delta\} \\ &\leq (h(\delta) + \delta \ln M) \Pr \{P_e(Y^\tau) \leq \delta\} + (\ln M) \Pr \{P_e(Y^\tau) > \delta\} \\ &\leq h(\delta) + \left( \delta + \frac{P_e}{\delta} \right) \ln M. \end{aligned}$$

This upper bound on the expected posterior entropy at time  $\tau$  can be turned into a lower bound on the expected value of  $\tau$  by using the channel capacity as an upper bound to the expected change of entropy. This notion is made precise by the following lemma,

*Lemma 2:* For any  $0 < \delta \leq \frac{1}{2}$

$$E[\tau] \geq \left( 1 - \delta - \frac{P_e}{\delta} \right) \frac{\ln M}{C} - \frac{h(\delta)}{C}.$$

*Proof:* Observe that  $\{\mathcal{H}(W|Y^n) + nC\}$  is a submartingale (an observation already made in [3, Lemma 2]). To see this,

$$\begin{aligned} E[\mathcal{H}(W|Y^n) - \mathcal{H}(W|Y^{n+1}) | Y^n = y^n] &= I(W; Y_{n+1} | Y^n = y^n) \\ &\stackrel{(a)}{\leq} I(X_{n+1}; Y_{n+1} | Y^n = y^n) \\ &\leq C \end{aligned}$$

where (a) follows from the data processing inequality and the fact that  $W - X_{n+1} - Y_{n+1}$  forms a Markov chain given  $Y^n = y^n$ . Hence  $\{\mathcal{H}(W|Y^n) + nC\}$  is indeed a submartingale. Since  $\mathcal{H}(W|y^n)$  is bounded between 0 and  $\ln M$  for all  $n$ , and the expected stopping time  $E[\tau] \leq E[T] < \infty$ , Doob's Optional-Stopping Theorem allows us to conclude that at time  $\tau$  the expected value of the submartingale must be greater than or equal to the initial value,  $\ln M$ . Hence

$$\begin{aligned} \ln M = \mathcal{H}(W|Y^0) &\leq E[\mathcal{H}(W|Y^\tau) + \tau C] \\ &= E[\mathcal{H}(W|Y^\tau)] + E[\tau] C \\ &\leq h(\delta) + \left( \delta + \frac{P_e}{\delta} \right) \ln M + E[\tau] C. \end{aligned}$$

Solving for  $E[\tau]$  yields

$$E[\tau] \geq \left( 1 - \delta - \frac{P_e}{\delta} \right) \frac{\ln M}{C} - \frac{h(\delta)}{C}.$$

■

<sup>5</sup>Notice that  $\mathcal{H}(W|y^n)$  is commonly written as  $H(W|Y^n = y^n)$ . We cannot use the standard notation since it becomes problematic when we substitute  $Y^n$  for  $y^n$  as we just did.

## V. BURNASHEV'S LOWER BOUND

In this section we will combine the two bounds we have established in the preceding sections to obtain a bound on the overall expected decoding time. Lemma 2 provides a lower bound on  $E[\tau]$  as a function of  $M$ ,  $\delta$  and  $P_e$ . We will show that a properly constructed binary hypothesis testing problem allows us to use Lemma 1 to lower bound the probability of error in terms of  $E[T - \tau | Y^\tau]$ . This in turn will lead us to the final bound on  $E[T]$ .

The next proposition states that a new channel output symbol can not change the *a posteriori* probability of any particular message by more than some constant factor when  $C_1$  is finite.

*Proposition 2:*  $C_1 < \infty$  implies

$$\lambda p(w|y^{n-1}) \leq p(w|y^n) \leq \frac{p(w|y^{n-1})}{\lambda},$$

where  $0 < \lambda = \min_{x,y} p(y|x) \leq \frac{1}{2}$ .

*Proof:* Using Bayes' rule, the posterior may be written recursively as

$$p(w|y^n) = p(w|y^{n-1}) \frac{p(y_n|x_n = f_n(w, y^{n-1}))}{p(y_n|y^{n-1})}.$$

The quotient may be upper and lower bounded using  $1 \geq p(y_n|x_n = f_n(w, y^{n-1})) \geq \lambda$  and  $1 \geq p(y_n|y^{n-1}) \geq \lambda$ , which yields the statement of the proposition.  $\blacksquare$

Our objective is to lower bound the probability of error of a decoder that decides at time  $T$ . The key idea is that a binary hypothesis decision such as deciding whether or not  $W$  lies in some set  $\mathcal{G}$  can be made at least as reliably as a decision on the value of  $W$  itself.

Given a set  $\mathcal{G}$  of messages, consider deciding between  $W \in \mathcal{G}$  and  $W \notin \mathcal{G}$  in the following way: given access to the original decoder's estimate  $\hat{W}$ , declare that  $W \in \mathcal{G}$  if  $\hat{W} \in \mathcal{G}$ , and declare  $W \notin \mathcal{G}$  otherwise. This binary decision is always correct when the original decoder's estimate  $\hat{W}$  is correct. Hence the probability of error of this (not necessarily optimal) binary decision rule cannot exceed the probability of error of the original decoder, for any set  $\mathcal{G}$ . Thus the error probability of the optimal decoder deciding at time  $T$  whether or not  $W \in \mathcal{G}$  is a lower bound to the error probability of any decoder that decodes  $W$  itself at time  $T$ . This fact is true even if the set  $\mathcal{G}$  is chosen at a particular stopping time  $\tau$  and the error probabilities we are calculating are conditioned on the observation  $Y^\tau$ .

For every realization of  $Y^\tau$ , the message set can be divided into two parts,  $\mathcal{G}(Y^\tau)$  and its complement  $\mathcal{W} \setminus \mathcal{G}(Y^\tau)$ , in such a way that both parts have an *a posteriori* probability greater than  $\lambda\delta$ . The rest of this paragraph describes how this is possible. From the definition of  $\tau$ , at time  $\tau - 1$  the *a posteriori* probability of every message is smaller than  $1 - \delta$ . This implies that the sum of the *a posteriori* probabilities of any set of  $M - 1$  messages is greater than  $\delta$  at time  $\tau - 1$ , and by Proposition 2, greater than  $\lambda\delta$  at time  $\tau$ . In particular,  $P_e(y^\tau) \geq \lambda\delta$ . We separately consider the cases  $P_e(y^\tau) \leq \delta$  and  $P_e(y^\tau) > \delta$ . In the first case,  $P_e(y^\tau) \leq \delta$ , let  $\mathcal{G}(Y^\tau)$  be the set consisting of only the message with the highest *a posteriori* probability at time  $\tau$ . The *a posteriori* probability of  $\mathcal{G}(Y^\tau)$  then satisfies  $\Pr\{\mathcal{G}(Y^\tau)\} \geq 1 - \delta \geq 1/2 \geq \lambda\delta$ . As argued above, its complement (the remaining  $M - 1$  messages) also has an *a posteriori* probability greater than  $\lambda\delta$ , thus for this  $\mathcal{G}(Y^\tau)$ ,  $\Pr\{\mathcal{G}(Y^\tau) | Y^\tau\} \in [\lambda\delta, 1 - \lambda\delta]$ . In the second case, namely when  $P_e(y^\tau) > \delta$ , the *a posteriori* probability of each message is smaller than  $1 - \delta$ . In this case the set  $\mathcal{G}(Y^\tau)$  may be formed by starting with the empty set and adding messages in arbitrary order until the threshold  $\delta/2$  is exceeded. This ensures that the *a priori* probability of  $\mathcal{G}(Y^\tau)$  is greater than  $\lambda\delta$ . Notice that the threshold will be exceeded by at most  $1 - \delta$ , thus the complement set has an *a posteriori* probability of at least  $\delta/2 > \lambda\delta$ . Thus  $\Pr\{\mathcal{G}(Y^\tau) | Y^\tau\} \in [\lambda\delta, 1 - \lambda\delta]$ .

For any realization of  $Y^\tau$  we have the binary hypothesis testing problem, running from  $\tau$  until  $T$ , deciding whether or not  $W \in \mathcal{G}(Y^\tau)$ . Notice that the *a priori* probabilities of the two hypotheses of this binary hypothesis

testing problem are the *a posteriori* probabilities of  $\mathcal{G}(Y^\tau)$  and  $\mathcal{W} \setminus \mathcal{G}(Y^\tau)$  at time  $\tau$  each of which is shown to be greater than  $\lambda\delta$  in the paragraph above. We apply Lemma 1 with  $A = \mathcal{G}(Y^\tau)$  and  $N = \mathcal{W} \setminus \mathcal{G}(Y^\tau)$  to lower bound the probability of error of the binary decision made at time  $T$  and, as argued above, we use the result to lower bound the probability that  $\hat{W} \neq W$ . Initially everything is conditioned on the channel output up to time  $\tau$ , thus

$$\Pr \left\{ \hat{W}(Y^T) \neq W \middle| Y^\tau \right\} \geq \frac{\lambda\delta}{4} e^{-C_1 E[T-\tau|Y^\tau]}.$$

Taking the expectation of the above expression over all realizations of  $Y^\tau$  yields the unconditional probability of error

$$P_e = E \left[ \Pr \left\{ \hat{W}(Y^T) \neq W \middle| Y^\tau \right\} \right] \geq E \left[ \frac{\lambda\delta}{4} e^{-C_1 E[T-\tau|Y^\tau]} \right].$$

Using the convexity of  $e^{-x}$  and Jensen's inequality, we obtain

$$P_e \geq \frac{\lambda\delta}{4} e^{-C_1 E[T-\tau]}.$$

Solving for  $E[T-\tau]$  yields

$$E[T-\tau] \geq \frac{-\ln P_e - \ln 4 + \ln(\lambda\delta)}{C_1}. \quad (17)$$

Combining Lemma 2 and (17) yields:

*Theorem 1:* The expected decoding time  $T$  of any variable-length block code for a DMC used with feedback is lower bounded by

$$E[T] \geq \left( 1 - \delta - \frac{P_e}{\delta} \right) \frac{\ln M}{C} + \frac{-\ln P_e}{C_1} - \frac{\ln(\delta)}{C} + \frac{\ln(\lambda\delta) - \ln 4}{C_1}, \quad (18)$$

where  $M$  is the cardinality of the message set,  $P_e$  the error probability,  $\lambda = \min_{x \in \mathcal{X}, y \in \mathcal{Y}} p(y|x)$ , and  $\delta$  is any number satisfying  $0 < \delta \leq \frac{1}{2}$ . ■

Choosing the parameter  $\delta$  as  $\delta = -\frac{1}{\ln P_e}$  achieves the required scaling for (5).

## VI. SUMMARY

We have presented a new derivation of Burnashev's asymptotically tight lower bound to the average delay needed for a target error probability when a message is communicated across a DMC used with (channel output) feedback. Our proof is simpler than the original, yet provides insight by clarifying the role played by the quantities that appear in the bound. Specifically, from the channel coding theorem we expect it to take roughly  $\frac{\ln M}{C}$  channel uses to reduce the probability of error of a MAP decision to some small (but not too small) value. At this point we can partition the message set in two subsets, such that neither subset has too small an *a posteriori* probability. From now on it takes (asymptotically)  $-\frac{\ln P_e}{C_1}$  channel uses to decide with probability of error  $P_e$  which of the two sets contains the true message. It takes at least as many channel uses to decide which message was selected and incur the same error probability.

For obvious reasons we may call the two phases the communication and the binary hypothesis testing phase, respectively. These two phases exhibit a pleasing similarity to the communication and confirmation phase of the optimal scheme proposed and analyzed by Yamamoto and Itoh in [4]. The fact that these two phases play a key role in proving achievability as well as in proving that one cannot do better suggests that they are an intrinsic component of an optimal communication scheme using variable-length block codes over DMCs with feedback.

## ACKNOWLEDGEMENT

The authors would like to thank R. G. Gallager for his help in pointing out an error in an earlier version of this paper and the reviewers for their helpful comments.

## REFERENCES

- [1] C. E. Shannon, "The zero error capacity of a noisy channel," *IEEE Trans. Inform. Theory*, vol. 2, no. 3, pp. 8–19, 1956.
- [2] I. Csiszár, "On the capacity of noisy channel with arbitrary signal costs," *Probl. Control and Inf. Theory*, vol. 2, pp. 283–304, 1973.
- [3] M. V. Burnashev, "Data transmission over a discrete channel with feedback," *Problemy Peredači Informacii*, vol. 12, no. 4, pp. 10–30, 1976. translated in Problems of Information Transmission, pp. 250–265, 1976.
- [4] H. Yamamoto and K. Itoh, "Asymptotic performance of a modified Schalkwijk-Barron scheme for channels with noiseless feedback," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 729–733, November 1979.
- [5] H. Chernoff, "Large-sample theory: Parametric case," *Ann. Math. Stat.*, vol. 27, pp. 1–22, March 1956.
- [6] G. Como, S. Yüksel, and S. Tatikonda, "On the error exponent of variable-length block-coding schemes over finite-state Markov channels with feedback," July 2007.
- [7] D. Williams, *Probability with Martingales*. Cambridge: Cambridge University Press, 1991.
- [8] M. Horstein, "Sequential transmission using noiseless feedback," *IEEE Trans. Inform. Theory*, vol. 9, no. 3, pp. 136–143, 1963.
- [9] B. D. Kudryashov, "Message transmission over a discrete channel with noiseless feedback," *Problemy Peredači Informacii*, vol. 15, no. 1, pp. 3–13, 1979. translated in Problems of Information Transmission, pp. 1–9, 1979.
- [10] A. Tchamkerten and I. E. Telatar, "On the universality of Burnashev's error exponent," *IEEE Trans. Inform. Theory*, vol. 51, no. 8, pp. 2940–2944, 2005.
- [11] I. Csiszár and P. Shields, "Information theory and statistics: A tutorial," in *Foundations and Trends in Communications and Information Theory*, vol. 1, now Publishers Inc, 2004.
- [12] T. M. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
- [13] I. Csiszár and J. Körner, *Information Theory*. New York: Academic Press, 1981.